Journal of Machine Learning Research 1 (2012) 1-48 Submitted 4/00; Published 10/00

Regret lower bounds and extended Upper Confidence Bounds policies in stochastic multi-armed bandit problem

Antoine Salomon SALOMONA@IMAGINE.ENPC.FR
Imagine, Université Paris Est

Jean-Yves Audibert AUDIBERT@IMAGINE.ENPC.FR
Imagine, Université Paris Est
&
Sierra, CNRS/ENS/INRIA, Paris, France

Issam El Alaoui ISSAM.EL-ALAOUI.2007@POLYTECHNIQUE.ORG
Imagine, Université Paris Est

Editor: ?

Abstract

This paper is devoted to regret lower bounds in the classical model of stochastic multi-armed bandit. A well-known result of Lai and Robbins, which has then been extended by Burnetas and Katehakis, has established the presence of a logarithmic bound for all consistent policies. We relax the notion of consistence, and exhibit a generalisation of the logarithmic bound. We also show the non existence of logarithmic bound in the general case of Hannan consistency. To get these results, we study variants of popular Upper Confidence Bounds (UCB) policies. As a by-product, we prove that it is impossible to design an adaptive policy that would select the best of two algorithms by taking advantage of the properties of the environment.

Keywords: stochastic bandits, regret bounds, selectivity, UCB policies.

1. Introduction and notations

Multi-armed bandits are a classical way to illustrate the difficulty of decision making in the case of a dilemma between exploration and exploitation. The denomination of these models comes from an analogy with playing a slot machine with more than one arm. Each arm has a given (and unknown) reward distribution and, for a given number of rounds, the agent has to choose one of them. As the goal is to maximize the sum of rewards, each round decision consists in a trade-off between exploitation (i.e. playing the arm that has been the more lucrative so far) and exploration (i.e. testing an other arm, hoping to discover an alternative that beats the current best choice). One possible application is clinical trial, when one wants to heal as many patients as possible, when the latter arrive sequentially and when the effectiveness of each treatment is initially unknown (Thompson, 1933). Bandit problems has initially been studied by Robbins (1952), and its interest has then been extended to many fields such as economics (Lamberton et al., 2004; Bergemann and Valimaki, 2008), games (Gelly and Wang, 2006), optimisation (Kleinberg, 2005; Coquelin and Munos, 2007;

Kleinberg et al., 2008; Bubeck et al., 2009),...

Let us detail our model. A stochastic multi-armed bandit problem is defined by:

- a number of rounds n,
- a number of arms $K \geq 2$,
- an environment $\theta = (\nu_1, \cdots, \nu_K)$, where each ν_k ($k \in \{1, \cdots, K\}$) is a real-valued measure that represents the distribution reward of arm k.

We assume that rewards are bounded. Thus, for simplicity, each ν_k is a probability on $[0, 1]$. Environment θ is initially unknown by the agent but lies in some known set Θ of the form $\Theta_1 \times \ldots \times \Theta_K$, meaning that Θ_k is the set of possible reward distributions of arm k. For the problem to be interesting, the agent should not have great knowledges of its environment, so that Θ should not be too small and/or contain too trivial distributions such as Dirac measures. To make it simple, each Θ_k is assumed to contain the distributions $p\delta_a + (1-p)\delta_b$, where $p, a, b \in [0, 1]$ and δ_x denotes the Dirac measure centred on x. In particular, it contains Dirac and Bernoulli distributions. Note that the number of rounds n may or may not be known by the agent, but this will not affect the present study. Some aspects of this particular point can be found in Salomon and Audibert (2011).
The game is as follows. At each round (or time step) $t = 1, \cdots, n$, the agent has to choose an arm I_t in the set of arms $\{1, \cdots, K\}$. This decision is based on past actions and observations, and the agent may also randomize his choice. Once the decision is made, the agent gets and observes a payoff that is drawn from ν_{I_t} independently from the past. Thus we can describe a policy (or strategy) as a sequence $(\sigma_t)_{t\geq 1}$ (or $(\sigma_t)_{1\leq t\leq n}$ if the number of rounds n is known) such that each σ_t is a mapping from the set $\{1, \ldots, K\}^{t-1} \times [0, 1]^{t-1}$ of past decisions and outcomes into the set of arm $\{1, \ldots, K\}$ (or into the set of probabilities on $\{1, \ldots, K\}$, in case the agent randomizes his choices).
For each arm k and all times t, let $T_k(t) = \sum_{s=1}^{t} \mathbb{1}_{I_s=k}$ denote the number of times arm k was pulled from round 1 to round t, and $X_{k,1}, X_{k,2}, \ldots, X_{k,T_k(t)}$ the corresponding sequence of rewards. We denote by $\mathbb{P}_\theta$ the distribution on the probability space such that for any $k \in \{1, \ldots, K\}$, the random variables $X_{k,1}, X_{k,2}, \ldots, X_{k,n}$ are i.i.d. realizations of ν_k, and such that these K sequences of random variables are independent. Let $\mathbb{E}_\theta$ denote the associated expectation.
Let $\mu_k = \int x d\nu_k(x)$ be the mean reward of arm k. Introduce $\mu^* = \max_{k\in\{1,\ldots,K\}} \mu_k$ and fix an arm $k^* \in \operatorname{argmax}_{k\in\{1,\ldots,K\}} \mu_k$, that is k^* has the best expected reward. The agent aims at minimizing its *regret*, defined as the difference between the cumulative reward he would have obtained by always drawing the best arm and the cumulative reward he actually received. Its regret is thus

$$R_n = \sum_{t=1}^{n} X_{k^*,t} - \sum_{t=1}^{n} X_{I_t,T_{I_t}(t)}.$$

As most of the publications on this topic, we focus on expected regret, for which one can check that:

$$\mathbb{E}_\theta R_n = \sum_{k=1}^{K} \Delta_k \mathbb{E}_\theta[T_k(n)], \tag{1}$$

where Δ_k is the *optimality gap* of arm k, defined by $\Delta_k = \mu^* - \mu_k$. We also define Δ as the gap between the best arm and the second best arm, i.e. $\Delta := \min_{k \neq k^*} \Delta_k$.

Previous works have shown the existence of lower bounds on the performance of a large class of policies. In this way Lai and Robbins (1985) proved a lower bound of the expected regret of order $\log n$ in a particular parametric framework, and they also exhibited optimal policies. This work has then been extended by Burnetas and Katehakis (1996). Both papers deal with *consistent* policies, meaning that all the policies considered are such that:

$$\forall a > 0,\ \forall \theta \in \Theta,\ \mathbb{E}_\theta[R_n] = o(n^a). \tag{2}$$

The logarithmic bound of Burnetas and Katehakis is expressed as follows. For all environment $\theta = (\nu_1, \cdots, \nu_K)$ and all $k \in \{1, \ldots, K\}$, let us set

$$D_k(\theta) := \inf_{\tilde{\nu}_k \in \Theta_k : \mathbb{E}[\tilde{\nu}_k] > \mu^*} KL(\nu_k, \tilde{\nu}_k),$$

where $KL(\nu, \mu)$ denotes the Kullback-Leibler divergence of measures ν and μ. Now fix a consistent policy and an environment $\theta \in \Theta$. If k is a suboptimal arm (i.e. $\mu_k \neq \mu^*$) such that $0 < D_k(\theta) < +\infty$, then

$$\forall \varepsilon > 0,\ \lim_{n \to +\infty} \mathbb{P}\left[T_k(n) \geq \frac{(1-\varepsilon)\log n}{D_k(\theta)} \right] = 1.$$

This readily implies that:

$$\liminf_{n \to +\infty} \frac{\mathbb{E}_\theta[T_k(n)]}{\log n} \geq \frac{1}{D_k(\theta)}.$$

Thanks to Equation (1), it is then easy to deduce a lower bound on the expected regret.
One contribution of this paper is to extend this bound to a larger class of policies. We will define the notion of α-consistency ($\alpha \in [0, 1]$) as a variant of Equation (2), where equality $\mathbb{E}_\theta[R_n] = o(n^a)$ only holds for all $a > \alpha$. We show that the logarithmic bound still holds, but coefficient $\frac{1}{D_k(\theta)}$ is turned into $\frac{1-\alpha}{D_k(\theta)}$. We also prove that the dependence of this new bound in the term $1 - \alpha$ is asymptotically optimal when $n \to +\infty$ (up to a constant).
As any policy achieves at most an expected regret of order n (because the average cost of pulling an arm k is a constant Δ_k), it is also natural to wonder what happens when expected regret is only required to be $o(n)$. This notion is equivalent to Hannan consistency. In this case, we show that there is no logarithmic bound any more.

Some of our results are obtained thanks to a study of particular Upper Confidence Bound algorithms. These policies were introduced by Lai and Robbins (1985): it basically consists in computing an index at each round and for each arm, and then in selecting the arm with the greatest index. A simple and efficient way to design such policies is to choose indexes that are upper bounds of the mean reward of the considered arm that hold with high probability (or, say, with high confidence level). This idea can be traced back to Agrawal (1995), and has been popularized by Auer et al. (2002), who notably described a policy called ucb1. For this policy, each index is defined by an arm k, a time step t, and

an integer s that indicates the number of times arm k has been pulled before stage t. It is denoted by $B_{k,s,t}$ and is given by:

$$B_{k,s,t} = \hat{X}_{k,s} + \sqrt{\frac{2\log t}{s}},$$

where $\hat{X}_{k,s}$ is the empirical mean of arm k after s pulls, i.e. $\hat{X}_{k,s} = \frac{1}{s}\sum_{u=1}^{s} X_{k,u}$.
To summarize, UCB1 policy first pulls each arm once and then, at each round $t > K$, selects an arm k that maximizes $B_{k,T_k(t-1),t}$. Note that, by means of Hoeffding's inequality, the index $B_{k,T_k(t-1),t}$ is indeed an upper bound of μ_k with high probability (i.e. the probability is greater than $1 - 1/t^4$). Note also that a way to look at this index is to interpret the empiric mean $\hat{X}_{k,T_k(t-1)}$ as an "exploitation" term, and the square root as an "exploration" term (as it gradually increases when arm k is not selected).
The policy UCB1 achieves the logarithmic bound (up to a multiplicative constant), as it was shown that:

$$\forall\theta \in \Theta,\ \forall n \geq 3,\quad \mathbb{E}_\theta[T_k(n)] \leq 12\frac{\log n}{\Delta_k^2} \quad \text{and} \quad \mathbb{E}_\theta R_n \leq 12\sum_{k=1}^{K}\frac{\log n}{\Delta_k} \leq 12\frac{\log n}{\Delta}.$$

Audibert et al. (2009) studied some variants of UCB1 policy. Among them, one consists in changing the $2\log t$ in the exploration term into $\rho \log t$, where $\rho > 0$. This can be interpreted as a way to tune exploration: the smaller ρ is, the better the policy will perform in simple environments where information is disclosed easily (for example when all reward distributions are Dirac measures). On the contrary, ρ has to be greater to face more challenging environments (typically when reward distributions are Bernoulli laws with close parameters).
This policy, that we denote UCB(ρ), was proven by Audibert et al. to achieve the logarithmic bound when $\rho > 1$, and the optimality was also obtained when $\rho > \frac{1}{2}$ for a variant of UCB(ρ). Bubeck (2010) showed in his PhD dissertation that their ideas actually enable to prove optimality of UCB(ρ) for $\rho > \frac{1}{2}$. Moreover, the case $\rho = \frac{1}{2}$ corresponds to a confidence level of $\frac{1}{t}$ (in view of Hoeffding's inequality, as above), and several studies (Lai and Robbins, 1985; Agrawal, 1995; Burnetas and Katehakis, 1996; Audibert et al., 2009; Honda and Takemura, 2010) have shown that this level is critical. We complete these works by a precise study of UCB(ρ) when $\rho \leq \frac{1}{2}$. We prove that UCB(ρ) is $(1-2\rho)$-consistent and that it is not α-consistent for any $\alpha < 1 - 2\rho$ (in view of the definition above, meaning that expected regret is roughly of order $n^{1-2\rho}$). Not surprisingly, it performs well in simple settings, represented by an environment where all reward distributions are Dirac measures. A by-product of this study is that it is not possible to design an algorithm that would specifically adapt to some kinds of environments, i.e. that would for example be able to select a proper policy depending on the environment being simple or challenging. In particular, and contrary to the results obtained within the class of consistent policies, there is no optimal policy. This contribution is linked with selectivity in on-line learning problem with perfect information, commonly addressed by prediction with expert advice such as algorithms with exponentially weighted forecasters (see, e.g., Cesa-Bianchi and Lugosi (2006)). In this spirit, a closely related problem to ours is the one of regret against the best strategy from a pool studied by Auer et al. (2003): the latter designed a policy in the

context of adversarial/nonstochastic bandit whose decisions are based on a given number of recommendations (experts), which are themselves possibly the rewards received by a set of given algorithms. To a larger extent, model selection have been intensively studied in statistics, and is commonly solved by penalization methods (Mallows, 1973; Akaike, 1973; Schwarz, 1978).
Finally, we exhibit expected regret lower bounds of more general ucb policies, with the $2\log t$ in the exploration term of ucb1 replaced by an arbitrary function. We obtain Hannan consistent policies and, as mentioned before, lower bounds need not be logarithmic any more.

The paper is organized as follows: in Section 2 we give bounds on the expected regret of ucb(ρ) ($\rho < \frac{1}{2}$). In Section 3 we study the problem of selectivity. Then we focus in Section 4 on α-consistent policies, and we conclude in Section 5 by results on Hannan consistency by means of extended ucb policies.
Throughout the paper $\lceil x \rceil$ denotes the smallest integer which greater than the real x, and $Ber(p)$ denotes the Bernoulli law with parameter p.

2. Bounds on the expected regret of ucb(ρ), $\rho < \frac{1}{2}$

In this section we study the performances of ucb(ρ) policy, with ρ lying in the interval $(0, \frac{1}{2})$. We recall that ucb(ρ) is defined by:

- Draw each arm once,
- Then at each round t, draw an arm

$$I_t \in \underset{k\in\{1,\dots,K\}}{\operatorname{argmax}} \left\{ \hat{X}_{k,T_k(t-1)} + \sqrt{\frac{\rho \log t}{T_k(t-1)}} \right\}.$$

Small values of ρ can be interpreted as a low level of experimentation in the balance between exploration and exploitation, and present literature has not provided precise regret bound orders of ucb(ρ) with $\rho \in (0, \frac{1}{2})$ yet.
We first study the policy in simple environments (i.e. all reward distributions are Dirac measures), where the policy is supposed to perform well. We show that its expected regret is of order $\frac{\rho \log n}{\Delta}$ (Proposition 1 for the upper bound and Proposition 2 for the lower bound). These good performances are compensated by poor results in complexer environments, as we then prove that the overall expected regret lower bound is roughly of order $n^{1-2\rho}$ (Theorem 3).

Proposition 1 *Let $0 \leq b < a \leq 1$ and $n \geq 1$. For $\theta = (\delta_a, \delta_b)$, the random variable $T_2(n)$ is uniformly upper bounded by $\frac{\rho}{\Delta^2}\log(n) + 1$. Consequently, the expected regret of ucb(ρ) is upper bounded by $\frac{\rho}{\Delta}\log(n) + 1$.*

Proof Let us prove the upper bound on the sampling time of the suboptimal arm by contradiction. The assertion is obviously true for $n = 1$ and $n = 2$. If the assertion is false, then there exists $t \geq 3$ such that $T_2(t) > \frac{\rho}{\Delta^2}\log(t) + 1$ and $T_2(t-1) \leq \frac{\rho}{\Delta^2}\log(t-1) + 1$.

Since $\log(t) \geq \log(t-1)$, this leads to $T_2(t) > T_2(t-1)$, meaning that arm 2 is drawn at time t. Therefore, we have $a + \sqrt{\frac{\rho \log(t)}{t-1-T_2(t-1)}} \leq b + \sqrt{\frac{\rho \log(t)}{T_2(t-1)}}$, hence $\Delta \leq \sqrt{\frac{\rho \log(t)}{T_2(t-1)}}$, which implies $T_2(t-1) \leq \frac{\rho \log(t)}{\Delta^2}$ and thus $T_2(t) \leq \frac{\rho \log(t)}{\Delta^2} + 1$. This contradicts the definition of t, which ends the proof of the first statement. The second statement is a direct consequence of Formula (1). ■

The following shows that Proposition 1 is tight and allows to conclude that the expected regret of UCB(ρ) is equivalent to $\frac{\rho}{\Delta} \log(n)$ when n goes to infinity.

Proposition 2 *Let $0 \leq b < a \leq 1$, $n \geq 2$ and $h : t \mapsto \frac{\rho}{\Delta^2} \log(t) \Big(1 + \sqrt{\frac{2\rho \log(t)}{(t-1)\Delta^2}}\Big)^{-2}$. For $\theta = (\delta_a, \delta_b)$, the random variable $T_2(n)$ is uniformly lower bounded by*

$$f(n) = \int_2^n \min\big(h'(s), 1\big) ds - h(2).$$

As a consequence, the expected regret of UCB(ρ) is lower bounded by $\Delta f(n)$.

Straightforward calculations shows that $h'(s) \leq 1$ for s large enough, and this explains why our lower bound $\Delta f(n)$ is equivalent to $\Delta h(n) \sim \frac{\rho}{\Delta} \log(n)$ as n goes to infinity.

Proof First, one can easily prove (for instance, by induction) that $T_2(t) \leq t/2$ for any $t \geq 2$. Let us prove the lower bound on $T_2(n)$ by contradiction. The assertion is obviously true for $n = 2$. If the assertion is false for $n \geq 3$, then there exists $t \geq 3$ such that $T_2(t) < f(t)$ and $T_2(t-1) \geq f(t-1)$. Since $f'(s) \in [0,1]$ for any $s \geq 2$, we have $f(t) \leq f(t-1) + 1$. These last three inequalities imply $T_2(t) < T_2(t-1) + 1$, which gives $T_2(t) = T_2(t-1)$. This means that arm 1 is drawn at time t. We consequently have

$$a + \sqrt{\frac{\rho \log(t)}{t-1-T_2(t-1)}} \geq b + \sqrt{\frac{\rho \log(t)}{T_2(t-1)}},$$

hence

$$\frac{\Delta}{\sqrt{\rho \log(t)}} \geq \frac{1}{\sqrt{T_2(t-1)}} - \frac{1}{\sqrt{t-1-T_2(t-1)}} \geq \frac{1}{\sqrt{T_2(t-1)}} - \frac{\sqrt{2}}{\sqrt{t-1}}.$$

We then deduce that $T_2(t) = T_2(t-1) \geq h(t) \geq f(t)$. This contradicts the definition of t, which ends the proof of the first statement. Again, the second statement results from Formula (1). ■

Now we show that the order of the lower bound of the expected regret is $n^{1-2\rho}$. Thus, for $\rho \in (0, \frac{1}{2})$, UCB($\rho$) does not perform enough exploration to achieve the logarithmic bound, as opposed to UCB(ρ) with $\rho \in (\frac{1}{2}, +\infty)$.

Theorem 3 *For any $\rho \in (0, \frac{1}{2})$, any $\theta \in \Theta$ and any $\beta \in (0,1)$, one has*

$$\mathbb{E}_\theta[R_n] \leq \sum_{k:\Delta_k>0} \frac{4\log n}{\Delta_k} + 2\Delta_k \left(\frac{\log n}{\log(1/\beta)} + 1\right) \frac{n^{1-2\rho\beta}}{1-2\rho\beta}.$$

Moreover, for any $\varepsilon > 0$, there exists $\theta \in \Theta$ such that

$$\lim_{n\to+\infty} \frac{\mathbb{E}_\theta[R_n]}{n^{1-2\rho-\varepsilon}} = +\infty.$$

Proof Let us first show the upper bound. The core of the proof is a peeling argument and makes use of Hoeffding's maximal inequality. The idea is originally taken from Audibert et al. (2009), and the following is an adaptation of the proof of an upper bound in the case $\rho > \frac{1}{2}$ which can be found in S. Bubeck's PhD dissertation.
First, let us notice that the policy selects arm k such that $\Delta_k > 0$ at step t only if at least one of the three following equations holds:

$$B_{k^*,T_{k^*}(t-1),t} \leq \mu^*, \tag{3}$$

$$\hat{X}_{k,t} \geq \mu_k + \sqrt{\frac{\rho\log t}{T_k(t-1)}}, \tag{4}$$

$$T_k(t-1) < \frac{4\rho\log n}{\Delta_k^2}. \tag{5}$$

Indeed, if none of the equations holds, then:

$$B_{k^*,T_{k^*}(t-1),t} > \mu^* = \mu_k + \Delta_k \geq \mu_k + 2\sqrt{\frac{\rho\log n}{T_k(t-1)}} > \hat{X}_{k,t} + \sqrt{\frac{\rho\log t}{T_k(t-1)}} = B_{k,T_k(t-1),t}.$$

We denote respectively by $\xi_{1,t}, \xi_{2,t}$ and $\xi_{3,t}$ the events corresponding to Equations (3), (4) and (5).
We have:

$$\begin{aligned}
\mathbb{E}_\theta[T_k(n)] &= \mathbb{E}\left[\sum_{t=1}^n \mathbb{1}_{I_t=k}\right] \leq \frac{4\log n}{\Delta_k^2} + \mathbb{E}\left[\sum_{t=\lceil 4\log n/\Delta_k^2\rceil}^n \mathbb{1}_{\{I_t=k\}\setminus\xi_{3,t}}\right] \\
&\leq \frac{4\log n}{\Delta_k^2} + \mathbb{E}\left[\sum_{t=\lceil 4\log n/\Delta_k^2\rceil}^n \mathbb{1}_{\xi_{1,t}\cup\xi_{2,t}}\right] \leq \frac{4\log n}{\Delta_k^2} + \sum_{t=\lceil 4\log n/\Delta_k^2\rceil}^n \mathbb{P}(\xi_{1,t}) + \mathbb{P}(\xi_{2,t}).
\end{aligned}$$

We now have to find a proper upper bound for $\mathbb{P}(\xi_{1,t})$ and $\mathbb{P}(\xi_{2,t})$. To this aim, we apply the peeling argument with a geometric grid over the time interval [1, t]:

$$\begin{aligned}
\mathbb{P}(\xi_{1,t}) &\leq \mathbb{P}\left(\exists s \in \{1,\cdots,t\},\ \hat{X}_{k^*,s} + \sqrt{\frac{\rho \log t}{s}} \leq \mu^*\right) \\
&\leq \sum_{j=0}^{\frac{\log t}{\log(1/\beta)}} \mathbb{P}\left(\exists s : \{\beta^{j+1}t < s \leq \beta^j t\},\ \hat{X}_{k^*,s} + \sqrt{\frac{\rho \log t}{s}} \leq \mu^*\right) \\
&\leq \sum_{j=0}^{\frac{\log t}{\log(1/\beta)}} \mathbb{P}\left(\exists s : \{\beta^{j+1}t < s \leq \beta^j t\},\ \sum_{l=1}^{s} X_{k^*,l} - \mu^* \leq -\sqrt{\rho s \log t}\right) \\
&\leq \sum_{j=0}^{\frac{\log t}{\log(1/\beta)}} \mathbb{P}\left(\exists s : \{\beta^{j+1}t < s \leq \beta^j t\},\ \sum_{l=1}^{s} X_{k^*,l} - \mu^* \leq -\sqrt{\rho \beta^{j+1} t \log t}\right).
\end{aligned}$$

By means of Hoeffding-Azumas inequality for martingales, we then have:

$$\mathbb{P}(\xi_{1,t}) \leq \sum_{j=0}^{\frac{\log t}{\log(1/\beta)}} \exp\left(-\frac{2\left(\sqrt{\beta^{j+1} t \rho \log t}\right)^2}{\beta^j t}\right) = \left(\frac{\log t}{\log(1/\beta)} + 1\right) \frac{1}{t^{2\rho\beta}},$$

and, for the same reasons, this bound also holds for $\mathbb{P}(\xi_{2,t})$.
Combining the former inequalities, we get:

$$\begin{aligned}
\mathbb{E}_\theta[T_k(n)] &\leq \frac{4\log n}{\Delta_k^2} + 2 \sum_{t=\lceil 4\log n/\Delta_k^2 \rceil}^{n} \left(\frac{\log t}{\log(1/\beta)} + 1\right) \frac{1}{t^{2\rho\beta}} \qquad (6) \\
&\leq \frac{4\log n}{\Delta_k^2} + 2\left(\frac{\log n}{\log(1/\beta)} + 1\right) \sum_{t=\lceil 4\log n/\Delta_k^2 \rceil}^{n} \frac{1}{t^{2\rho\beta}} \\
&\leq \frac{4\log n}{\Delta_k^2} + 2\left(\frac{\log n}{\log(1/\beta)} + 1\right) \int_1^n \frac{1}{t^{2\rho\beta}} dt \\
&\leq \frac{4\log n}{\Delta_k^2} + 2\left(\frac{\log n}{\log(1/\beta)} + 1\right) \frac{n^{1-2\rho\beta}}{1-2\rho\beta}.
\end{aligned}$$

As usual, the bound on the expected regret then comes formula (1).

Now let us show the lower bound. The result is obtained by considering an environment θ of the form $\left(Ber(\frac{1}{2}), \delta_{\frac{1}{2}-\Delta}\right)$, where $\Delta > 0$ is such that $2\rho(1+\sqrt{\Delta})^2 < 2\rho + \varepsilon$. We set $T_n := \lceil \frac{\rho \log n}{\Delta} \rceil$, and define the event ξ_n by:

$$\xi_n = \left\{\hat{X}_{1,T_n} < \frac{1}{2} - (1 + \frac{1}{\sqrt{\Delta}})\Delta\right\}.$$

When event ξ_n occurs, for any $t \in \{T_n, \dots, n\}$ one has

$$\begin{aligned} \hat{X}_{1,T_n} + \sqrt{\frac{\rho \log t}{T_n}} &\leq \hat{X}_{1,T_n} + \sqrt{\frac{\rho \log n}{T_n}} < \frac{1}{2} - (1 + \frac{1}{\sqrt{\Delta}})\Delta + \sqrt{\Delta} \\ &\leq \frac{1}{2} - \Delta, \end{aligned}$$

so that arm 1 is chosen no more than T_n times by ucb(ρ) policy. Thus:

$$\mathbb{E}_\theta [T_2(n)] \geq \mathbb{P}_\theta(\xi_n)(n - T_n).$$

We shall now find a lower bound of the probability of ξ_n thanks to Berry-Esseen inequality. We denote by C the corresponding constant, and by Φ the c.d.f. of the standard normal distribution. For convenience, we also define the following quantities:

$$\sigma := \sqrt{\mathbb{E}\left[\left(X_{1,1} - \frac{1}{2}\right)^2\right]} = \frac{1}{2}, \; M_3 := \mathbb{E}\left[\left|X_{1,1} - \frac{1}{2}\right|^3\right] = \frac{1}{8}.$$

Using the fact that $\Phi(-x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}x}\beta(x)$ with $\beta(x) \xrightarrow[x \to +\infty]{} 1$, we are then able to write:

$$\begin{aligned} \mathbb{P}_\theta(\xi_n) &= \mathbb{P}_\theta\left(\frac{\hat{X}_{1,T_n} - \frac{1}{2}}{\sigma}\sqrt{T_n} \leq -2\left(1 + \frac{1}{\sqrt{\Delta}}\right)\Delta\sqrt{T_n}\right) \\ &\geq \Phi\left(-2(\Delta + \sqrt{\Delta})\sqrt{T_n}\right) - \frac{CM_3}{\sigma^3\sqrt{T_n}} \\ &\geq \frac{\exp\left(-2(\frac{\rho \log n}{\Delta} + 1)(\Delta + \sqrt{\Delta})^2\right)}{2\sqrt{2\pi}(\Delta + \sqrt{\Delta})\sqrt{T_n}}\beta\left(2(\Delta + \sqrt{\Delta})\sqrt{T_n}\right) - \frac{CM_3}{\sigma^3\sqrt{T_n}} \\ &\geq n^{-2\rho(1+\sqrt{\Delta})^2}\frac{\exp\left(-2(\Delta + \sqrt{\Delta})^2\right)}{2\sqrt{2\pi}(\Delta + \sqrt{\Delta})\sqrt{T_n}}\beta\left(2(\Delta + \sqrt{\Delta})\sqrt{T_n}\right) - \frac{CM_3}{\sigma^3\sqrt{T_n}}. \end{aligned}$$

Previous calculations and Formula (1) gives

$$\mathbb{E}_\theta[R_n] = \Delta\mathbb{E}_\theta[T_2(n)] \geq \Delta\mathbb{P}_\theta(\xi_n)(n - T_n)$$

and the former inequality easily leads to the conclusion of the theorem. ■

3. Selectivity

In this section, we address the problem of selectivity in multi-armed stochastic bandit models. By selectivity, we mean the ability to adapt to the environment as and when rewards are observed. More precisely, it refers to the existence of a procedure that would perform at least as good as the policy that is best suited to the current environment θ among a given set of two (or more) policies. Two mains reasons motivates this study.
On the one hand this question was answered by Burnetas and Katehakis within the class of consistent policies. Let us recall the definition of consistent policies.

Definition 4 *A policy is consistent if*

$$\forall a > 0,\ \forall \theta \in \Theta,\ \mathbb{E}_\theta[R_n] = o(n^a).$$

Indeed they show the existence of lower bounds on the expected regret (see Section 3, Theorem 1 of Burnetas and Katehakis (1996)), which we also recall for the sake of completeness.

Theorem 5 *Fix a consistent policy and $\theta \in \Theta$. If k is a suboptimal arm (i.e. $\mu_k < \mu^*$) and if $0 < D_k(\theta) < +\infty$, then*

$$\forall \varepsilon > 0,\ \lim_{n\to+\infty} \mathbb{P}_\theta \left[T_k(n) \geq \frac{(1-\varepsilon)\log n}{D_k(\theta)} \right] = 1.$$

Consequently

$$\liminf_{n\to+\infty} \frac{\mathbb{E}_\theta[T_k(n)]}{\log n} \geq \frac{1}{D_k(\theta)}.$$

Remind that the lower bound on the expected regret is then deduced from formula (1).
Burnetas and Katehakis then exhibits an asymptotically optimal policy, i.e. which achieves the former lower bounds. The fact that a policy does as best as any other one obviously solves the problem of selectivity.
Nevertheless one can wonder what happens if we do not restrict our attention to consistent policies any more. Thus, one natural way to relax the notion of consistency is the following.

Definition 6 *A policy is α-consistent if*

$$\forall a > \alpha,\ \forall \theta \in \Theta,\ \mathbb{E}_\theta[R_n] = o(n^a).$$

For example we showed in the former section that UCB(ρ) is $(1-2\rho)$-consistent for any $\rho \in (0, \frac{1}{2})$. The class of α-consistent policies will be studied in Section 4.
Moreover, as the expected regret of any policy is at most of order n, it seems simpler and relevant to only require it to be $o(n)$:

$$\forall \theta \in \Theta,\ \mathbb{E}_\theta[R_n] = o(n),$$

which corresponds to the definition of Hannan consistency. The class of Hannan consistent policies includes consistent policies and α-consistent policies for any $\alpha \in (0,1)$. Some results on Hannan consistency will be provided in Section 5.

On the other hand, this problem has already been studied in the context of adversarial bandit by Auer et al. (2003). Their setting differs from our not only because their bandits are nonstochastic, but also because their adaptive procedure takes only into account a given number of recommendations, whereas in our setting the adaptation is supposed to come from observing rewards of the chosen arms (only one per time step). Nevertheless, there are no restrictions about consistency in the adversarial context and one can wonder if an "exponentially weighted forecasters" procedure like Exp4 could be transposed to our context. The answer is negative, as stated in the following theorem.

Theorem 7 *Let $\tilde{\mathcal{A}}$ be a consistent policy and let ρ be a real in $(0, 0.4)$. There are no policy which can both beat $\tilde{\mathcal{A}}$ and* ucb*(ρ), i.e.:*

$$\forall A,\ \exists \theta \in \Theta,\ \limsup_{n\to+\infty} \frac{\mathbb{E}_\theta[R_n(A)]}{\min(\mathbb{E}_\theta[R_n(\tilde{\mathcal{A}})], \mathbb{E}_\theta[R_n(\text{UCB}(\rho))])} > 1.$$

Thus there are no optimal policy if we extend the notion of consistency. Precisely, as ucb(ρ) is $(1-2\rho)$-consistent, we have shown that there are no optimal policy within the class of α-consistent policies (which is included in the class of Hannan consistent policies), where $\alpha > 0.2$.
Moreover, ideas from selectivity in adversarial bandits can not work in the present context. As we said, this impossibility may also come from the fact that we can not observe at each step the decisions and rewards of more than one algorithm. Nevertheless, if we were able to observe a given set policies from step to step, then it would be easy to beat them all: it is then sufficient to aggregate all the observations and simply pull the arm with the greater empiric mean. The case where we only observe decisions (and not rewards) of a set of policies may be interesting, but is left outside of the scope of this paper.

Proof Assume by contradiction that

$$\exists A,\ \forall \theta \in \Theta,\ \limsup_{n\to+\infty} u_{n,\theta} \leq 1,$$

where $u_{n,\theta} = \frac{\mathbb{E}_\theta[R_n(A)]}{\min(\mathbb{E}_\theta[R_n(\tilde{\mathcal{A}})], \mathbb{E}_\theta[R_n(UCB(\rho))])}$.
One has

$$\mathbb{E}_\theta[R_n(A)] \leq u_{n,\theta}\mathbb{E}_\theta[R_n(\tilde{\mathcal{A}})],$$

so that the fact that $\tilde{\mathcal{A}}$ is a consistent policy implies that A is also consistent. Consequently the lower bound of Burnetas and Katehakis has to hold. In particular, in environment $\theta = (\delta_0, \delta_\Delta)$ one has for any $\varepsilon > 0$ and with positive probability (provided that n is large enough):

$$T_1(n) \geq \frac{(1-\varepsilon)\log n}{D_k(\theta)}.$$

Now, note that there is simple upper bound of $D_k(\theta)$:

$$\begin{aligned} D_k(\theta) &\leq \inf_{p,a\in[0,1]:(1-p)a>\Delta} KL(\delta_0, p\delta_0 + (1-p)\delta_a) \\ &= \inf_{p,a\in[0,1]:(1-p)a>\Delta} \log\left(\frac{1}{p}\right) = \log\left(\frac{1}{1-\Delta}\right). \end{aligned}$$

And on the other hand, one has by means of Proposition 2:

$$T_1(n) \leq 1 + \frac{\rho\log n}{\Delta^2}.$$

Thus we have that, for any $\varepsilon > 0$ and if n is large enough

$$1 + \frac{\rho\log n}{\Delta^2} \geq \frac{(1-\varepsilon)\log n}{\log\left(\frac{1}{1-\Delta}\right)}$$

Letting ε go to zero and n to infinity, we get:

$$\frac{\rho}{\Delta^2} \geq \frac{1}{\log\left(\frac{1}{1-\Delta}\right)}.$$

This means that ρ has to be lower bounded by $\frac{\Delta^2}{\log\left(\frac{1}{1-\Delta}\right)}$, but this is greater than 0.4 if $\Delta = 0.75$, hence the contradiction.

■

Note that the former proof give us a simple alternative to Theorem 3 to show that UCB(ρ) is not consistent if $\rho \leq 0.4$. Indeed if it were consistent, then in environment $\theta = (\delta_0, \delta_\Delta)$, $T_1(n)$ would also have to be greater than $\frac{(1-\varepsilon)\log n}{D_k(\theta)}$ and lower than $1 + \frac{\rho \log n}{\Delta^2}$, and the same contradiction would hold.

4. Bounds on α-consistent policies

We now study α-consistent policies. We first show that the main result of Burnetas and Katehakis (Theorem 5) can be extended in the following way.

Theorem 8 *Fix an α-consistent policy and $\theta \in \Theta$. If k is a suboptimal arm and if $0 < D_k(\theta) < +\infty$, then*

$$\forall \varepsilon > 0, \ \lim_{n\to+\infty} \mathbb{P}_\theta \left[T_k(n) \geq (1-\varepsilon)\frac{(1-\alpha)\log n}{D_k(\theta)} \right] = 1.$$

Consequently

$$\liminf_{n\to+\infty} \frac{\mathbb{E}_\theta[T_k(n)]}{\log n} \geq \frac{1-\alpha}{D_k(\theta)}.$$

Recall that, as opposed to Burnetas and Katehakis (1996), there are no optimal policy (i.e. a policy that would achieve the lower bounds in all environment θ), as proven in the former section.

Proof We adapt Proposition 1 in Burnetas and Katehakis (1996) and its proof, which one may have a look at for further details. We fix $\varepsilon > 0$, and we want to show that:

$$\lim_{n\to+\infty} \mathbb{P}_\theta \left(\frac{T_k(n)}{\log n} \geq \frac{(1-\varepsilon)(1-\alpha)}{D_k(\theta)} \right) = 0.$$

Set $\delta > 0$ and $\delta' > \alpha$ such that $\frac{1-\delta'}{1+\delta} > (1-\varepsilon)(1-\alpha)$. By definition of $D_k(\theta)$, there exists $\tilde{\theta}$ such that $\mathbb{E}_{\tilde{\theta}}[X_{k,1}] > \mu^*$ and

$$D_k(\theta) < KL(\nu_k, \tilde{\nu}_k) < (1+\delta)D_k(\theta),^1$$

1. In Burnetas and Katehakis (1996), $D_k(\theta)$ is denoted $\mathbf{K}_a(\underline{\theta})$ and $KL(\nu_k, \tilde{\nu}_k)$ is denoted $\mathbf{I}(\underline{\theta}_\mathbf{a}, \underline{\theta}'_\mathbf{a})$. The equivalence between other notations is straightforward.

where we denote $\theta = (\nu_1, \dots, \nu_K)$ and $\tilde{\theta} = (\tilde{\nu}_1, \dots, \tilde{\nu}_K)$.
Let us define $I^\delta = KL(\nu_k, \tilde{\nu}_k)$ and the sets

$$A_n^{\delta'} := \left\{ \frac{T_k(n)}{\log n} < \frac{1-\delta'}{I^\delta} \right\}, \; C_n^{\delta''} := \left\{ \log L_{T_k(n)} \leq \left(1-\delta''\right) \log n \right\},$$

where δ'' is such that $\alpha < \delta'' < \delta'$ and L_j is defined by $\log L_j = \sum_{i=1}^j \log\left(\frac{d\nu_k}{d\tilde{\nu}_k}(X_{k,i})\right)$.
We show that $\mathbb{P}_\theta(A_n^{\delta'}) = \mathbb{P}_\theta(A_n^{\delta'} \cap C_n^{\delta''}) + \mathbb{P}_\theta(A_n^{\delta'} \setminus C_n^{\delta''}) \xrightarrow[n\to+\infty]{} 0$.
On the one hand, one has:

$$\begin{aligned}
\mathbb{P}_\theta(A_n^{\delta'} \cap C_n^{\delta''}) &\leq e^{(1-\delta'')\log n} \mathbb{P}_{\tilde{\theta}}(A_n^{\delta'} \cap C_n^{\delta''}) && (7)\\
&\leq n^{1-\delta''} \mathbb{P}_{\tilde{\theta}}(A_n^{\delta'}) = n^{1-\delta''} \mathbb{P}_{\tilde{\theta}}\left(n - T_k(n) > n - \frac{1-\delta'}{I^\delta} \log n \right) \\
&\leq \frac{n^{1-\delta''} \mathbb{E}_{\tilde{\theta}}[n - T_k(n)]}{n - \frac{1-\delta'}{I^\delta} \log n} && (8)\\
&\leq \frac{\sum_{l \neq k} n^{-\delta''} \mathbb{E}_{\tilde{\theta}}[T_l(n)]}{1 - \frac{1-\delta'}{I^\delta} \frac{\log n}{n}} \xrightarrow[n\to+\infty]{} 0,
\end{aligned}$$

where (7) is consequence of the definition of $C_n^{\delta''}$, (8) comes from Markov's inequality, and where the final limit is a consequence of the α-consistence.
On the other hand we set $b_n := \frac{1-\delta'}{I^\delta} \log n$, so that we have:

$$\begin{aligned}
\mathbb{P}_\theta(A_n^{\delta'} \setminus C_n^{\delta''}) &\leq \mathbb{P}\left(\max_{j \leq \lfloor b_n \rfloor} \log L_j > (1-\delta'') \log n \right) \\
&\leq \mathbb{P}\left(\frac{1}{b_n} \max_{j \leq \lfloor b_n \rfloor} \log L_j > I^\delta \frac{1-\delta''}{1-\delta'} \right).
\end{aligned}$$

This term then tends to zero, as a consequence of the law of large numbers.
Now that $\mathbb{P}_\theta(A_n^{\delta'})$ tends to zero, the conclusion comes from the following inequality:

$$\frac{1-\delta'}{I^\delta} > \frac{1-\delta'}{(1+\delta) D_k(\theta)} \geq \frac{(1-\varepsilon)(1-\alpha)}{D_k(\theta)}.$$

■

The former lower bound is asymptotically optimal, as claimed in the following proposition.

Proposition 9 *There exists $\theta \in \Theta$ and a constant $c > 0$ such that, for any $\alpha \in [0,1)$, there exists an α-consistent policy and $k \neq k^*$ such that:*

$$\liminf_{n\to+\infty} \frac{\mathbb{E}_\theta[T_k(n)]}{(1-\alpha)\log n} \leq c.$$

Proof By means of Proposition 1, the following holds for UCB(ρ) in any environment of the form $\theta = (\delta_a, \delta_b)$ with $a \neq b$:

$$\liminf_{n\to+\infty} \frac{\mathbb{E}_\theta T_k(n)}{\log n} \leq \frac{\rho}{\Delta^2},$$

where $k \neq k^*$.
As UCB(ρ) is $(1-2\rho)$-consistent (Theorem 3), we can conclude by setting $c = \frac{1}{2\Delta^2}$ and by choosing the policy UCB($\frac{1-\alpha}{2}$).

■

5. Hannan consistency and other exploration functions

We now study the class of Hannan consistent policies. We first show the necessity to have a logarithmic lower bound in some environments θ, and then a study of extended UCB policies will prove that there does not exist a logarithmic bound on the whole set Θ.

5.1 The necessity of a logarithmic regret in some environments

A simple idea enables to understand the necessity of a logarithmic regret in some environments. Assume that the agent knows the number of rounds n, and that he balances exploration and exploitation in the following way: he first pulls each arm $s(n)$ times, and then selects the arm that has obtained the best empiric mean for the rest of the game. If we denote by $p_{s(n)}$ the probability that the best arm does not have the best empiric mean after the exploration phase (i.e. after the first $Ks(n)$ rounds), then the expected regret is of the form

$$c_1(1-p_{s(n)})s(n) + c_2 p_{s(n)} n. \tag{9}$$

Indeed if the agent manages to match the best arm then he only suffers the pulls of suboptimal arms during the exploration phase, and that represents an expected regret of order $s(n)$. If not, the number of pulls of suboptimal arms is of order n, and so is the expected regret.
Now we can approximate $p_{s(n)}$, because it has the same order as the probability that the best arm gets an empiric mean lower than the second best mean reward, and because $\frac{X_{k^*,s(n)}-\mu^*}{\sigma}\sqrt{s(n)}$ (where σ is the variance of $X_{k^*,1}$) approximately has a standard normal distribution by the central limit theorem:

$$\begin{aligned}
p_{s(n)} &\approx \mathbb{P}_\theta(X_{k^*,s(n)} \leq \mu^* - \Delta) = \mathbb{P}_\theta\left(\frac{X_{k^*,s(n)}-\mu^*}{\sigma}\sqrt{s(n)} \leq -\frac{\Delta\sqrt{s(n)}}{\sigma}\right) \\
&\approx \frac{1}{\sqrt{2\pi}}\frac{\sigma}{\Delta\sqrt{s(n)}}\exp\left(-\frac{1}{2}\left(\frac{\Delta\sqrt{s(n)}}{\sigma}\right)^2\right) \\
&\approx \frac{1}{\sqrt{2\pi}}\frac{\sigma}{\Delta\sqrt{s(n)}}\exp\left(-\frac{\Delta^2 s(n)}{2\sigma^2}\right).
\end{aligned}$$

Then it is clear why the expected regret has to be logarithmic: $s(n)$ has to be greater than $\log n$ if we want the second term $p_{s(n)}n$ of Equation (9) to be sub-logarithmic, but then first term $(1-p_{s(n)})s(n)$ is greater than $\log n$.

This idea can be generalized, and this gives the following proposition.

Proposition 10 *For any policy, there exists $\theta \in \Theta$ and such that*

$$\limsup_{n\to+\infty} \frac{\mathbb{E}_\theta R_n}{\log n} > 0.$$

This result can be seen as a consequence of the main result of Burnetas and Katehakis (Theorem 5): if we assume by contradiction that $\limsup_{n\to+\infty} \frac{\mathbb{E}_\theta R_n}{\log n} = 0$ for all θ, the considered policy is therefore consistent, but then the logarithmic lower bounds have to hold. The reason why we wrote the proposition anyway is that our proof is based on the simple reasoning stated above and that it consequently holds beyond our model (see the following for details).

Proof The proposition results from the following property on Θ: there exists two environments $\theta = (\nu_1, \dots, \nu_K)$ and $\tilde{\theta} = (\tilde{\nu}_1, \dots, \tilde{\nu}_K)$ and $k \in \{1, \dots, K\}$ such that

- k has the best mean reward in environment θ,
- k is not the winning arm in environment $\tilde{\theta}$,
- $\nu_k = \tilde{\nu}_k$ and there exists $\eta \in (0,1)$ such that

$$\prod_{\ell \neq k} \frac{d\nu_\ell}{d\tilde{\nu}_\ell}(X_{\ell,1}) \geq \eta \quad \mathbb{P}_{\tilde{\theta}} - a.s. \tag{10}$$

The idea is the following: in case $\nu_k = \tilde{\nu}_k$ is likely to be the reward distribution of arm k, then arm k has to be pulled often for the regret to be small if the environment is θ, but not so much, as one has to explore to know if the environment is actually $\tilde{\theta}$ (and the third condition ensures that the distinction can be tough to make). The lower bound on exploration is of order $\log n$, as in the sketch in the beginning of the section.

The proof actually holds for any Θ that has the above-mentioned property (i.e. without the assumptions we made on Θ, i.e. being of the form $\Theta_1 \times \ldots \times \Theta_K$ and/or containing distributions of the form $p\delta_a + (1-p)\delta_b$). In our setting, the property is easy to check. Indeed the three conditions hold for any k and any pair of environments $\theta = (\nu_1, \dots, \nu_K)$, $\tilde{\theta} = (\tilde{\nu}_1, \dots, \tilde{\nu}_K)$ such that each ν_ℓ (resp. $\tilde{\nu}_\ell$) is a Bernoulli law with parameter p_ℓ (resp. $\tilde{p}_\ell$) and such that

- $\forall \ell \neq k,\ \tilde{p}_k > \tilde{p}_\ell$,
- $\exists \ell \neq k,\ p_k < p_\ell$,
- $\tilde{p}_k = p_k$ and $p_\ell,\ \tilde{p}_\ell \in (0,1)$ for any $\ell \neq k$.

It is then sufficient to set

$$\eta = \left(\min \left\{ \frac{p_1}{\tilde{p}_1}, \dots, \frac{p_{k-1}}{\tilde{p}_{k-1}}, \frac{p_{k+1}}{\tilde{p}_{k+1}}, \dots, \frac{p_K}{\tilde{p}_K}, \frac{1-p_1}{1-\tilde{p}_1}, \dots, \frac{1-p_{k-1}}{1-\tilde{p}_{k-1}}, \frac{1-p_{k+1}}{1-\tilde{p}_{k+1}}, \dots, \frac{1-p_K}{1-\tilde{p}_K} \right\} \right)^{K-1},$$

as $\frac{d\nu_\ell}{d\tilde{\nu}_\ell}(X_{\ell,1})$ equals $\frac{p_\ell}{\tilde{p}_\ell}$ when $X_{\ell,1} = 1$ and $\frac{1-p_\ell}{1-\tilde{p}_\ell}$ when $X_{\ell,1} = 0$.

We will now compute a lower bound of the expected regret in environment $\tilde{\theta}$. To this aim, we set

$$g(n) := \frac{2\mathbb{E}_\theta R_n}{\Delta}.$$

In the following, $\tilde{\Delta}_k$ denotes the optimality gap of arm k in environment $\tilde{\theta}$. Moreover the switch from $\tilde{\theta}$ to θ will result from Equality (10) and from the fact that event $\left\{ \sum_{\ell \neq k} T_\ell(n) \leq g(n) \right\}$ is measurable with respect to $X_{\ell,1}, \dots, X_{\ell,\lfloor g(n) \rfloor}$ $(\ell \neq k)$ and to $X_{k,1}, \dots, X_{k,n}$. That enables us to introduce the function q such that

$$\mathbb{1}_{\left\{ \sum_{\ell \neq k} T_\ell(n) \leq g(n) \right\}} = q\big((X_{k,s})_{s=1..n}, (X_{\ell,s})_{\ell \neq k,\ s=1..\lfloor g(n) \rfloor} \big)$$

and to write:

$$\begin{aligned}
\mathbb{E}_{\tilde{\theta}} R_n &\geq \tilde{\Delta}_k \mathbb{E}_{\tilde{\theta}}[T_k(n)] \geq \tilde{\Delta}_k (n - g(n)) \mathbb{P}_{\tilde{\theta}} \left(T_k(n) \geq n - g(n) \right) \\
&= \tilde{\Delta}_k (n - g(n)) \mathbb{P}_{\tilde{\theta}} \left(\sum_{\ell \neq k} T_\ell(n) \leq g(n) \right) \\
&= \tilde{\Delta}_k (n - g(n)) \int q\big((x_{\ell,s})_{\ell \neq k,\ s=1..t}, (x_{k,s})_{s=1..n} \big) \prod_{\substack{\ell \neq k \\ s=1..\lfloor g(n) \rfloor}} d\tilde{\nu}_\ell(x_{\ell,s}) \prod_{s=1..n} d\tilde{\nu}_k(x_{k,s}) \\
&\geq \tilde{\Delta}_k (n - g(n)) \eta^{\lfloor g(n) \rfloor} \int q\big((x_{\ell,s})_{\ell \neq k,\ s=1..t}, (x_{k,s})_{s=1..n} \big) \prod_{\substack{\ell \neq k \\ s=1..\lfloor g(n) \rfloor}} d\nu_\ell(x_{\ell,s}) \prod_{s=1..n} d\nu_k(x_{k,s}) \\
&\geq \tilde{\Delta}_k (n - g(n)) \eta^{g(n)} \mathbb{P}_\theta \left(\sum_{\ell \neq k} T_\ell(n) \leq g(n) \right) \\
&= \tilde{\Delta}_k (n - g(n)) \eta^{g(n)} \left(1 - \mathbb{P}_\theta \left(\sum_{\ell \neq k} T_\ell(n) > g(n) \right) \right) \\
&\geq \tilde{\Delta}_k (n - g(n)) \eta^{g(n)} \left(1 - \frac{\mathbb{E}_\theta \left(\sum_{\ell \neq k} T_\ell(n) \right)}{g(n)} \right) \\
&\geq \tilde{\Delta}_k (n - g(n)) \eta^{g(n)} \left(1 - \frac{\mathbb{E}_\theta \left(\sum_{\ell \neq k} \Delta_\ell T_\ell(n) \right)}{\Delta g(n)} \right) \\
&\geq \tilde{\Delta}_k (n - g(n)) \eta^{g(n)} \left(1 - \frac{\mathbb{E}_\theta R_n}{\Delta g(n)} \right) = \tilde{\Delta}_k \frac{n - g(n)}{2} \eta^{g(n)},
\end{aligned}$$

where the very first inequality is a consequence of Formula (1).
We are now able to conclude. Indeed, if we assume that $\frac{\mathbb{E}_\theta R_n}{\log n} \xrightarrow[n\to+\infty]{} 0$, then one has $g(n) \leq \min\left(\frac{n}{2}, \frac{-\log n}{2\log \eta}\right)$ for n large enough and:

$$\mathbb{E}_{\tilde{\theta}} R_n \geq \tilde{\Delta}_k \frac{n-g(n)}{2} \eta^{g(n)} \geq \tilde{\Delta}_k \frac{n}{4} \eta^{\frac{-\log n}{2\log \eta}} = \tilde{\Delta}_k \frac{\sqrt{n}}{4}.$$

In particular, we have $\frac{\mathbb{E}_{\tilde{\theta}} R_n}{\log n} \xrightarrow[n\to+\infty]{} +\infty$, hence the conclusion.

■

To finish this section, note that a proof could have been written in the same way with a slightly different property on Θ: there exists two environments $\theta = (\nu_1, \ldots, \nu_K)$ and $\tilde{\theta} = (\tilde{\nu}_1, \ldots, \tilde{\nu}_K)$ and $k \in \{1, \ldots, K\}$ such that

- k has the best mean reward in environment θ,
- k is not the winning arm in environment $\tilde{\theta}$,
- $\nu_\ell = \tilde{\nu}_\ell$ for all $\ell \neq k$ and there exists $\eta \in (0,1)$ such that

$$\frac{d\nu_k}{d\tilde{\nu}_k}(X_{k,1}) \geq \eta \quad \mathbb{P}_{\tilde{\theta}} - a.s.$$

The dilemma is then between exploring arm k or pulling the best arm of environment $\tilde{\theta}$.

5.2 There are no logarithmic bound in general

We extend our study to more general ucb policies, and we will find that there does not exist logarithmic lower bounds of the expected regret in the case of Hannan consistency. With "ucb", we now refer to an ucb policy with indexes of the form:

$$B_{k,s,t} = \hat{X}_{k,s} + \sqrt{\frac{f_k(t)}{s}}$$

where functions f_k $(1 \leq k \leq K)$ are increasing.

To find conditions for Hannan consistency, let us first show the following upper bound.

Lemma 11 *If arm k does not have the best mean reward, then for any $\beta \in (0,1)$ the following upper bound holds:*

$$\mathbb{E}_\theta[T_k(n)] \leq u + \sum_{t=u+1}^{n} \left(1 + \frac{\log t}{\log(\frac{1}{\beta})}\right) \left(e^{-2\beta f_k(t)} + e^{-2\beta f_{k^*}(t)}\right),$$

where $u = \left\lceil \frac{4 f_k(n)}{\Delta_k} \right\rceil$.

Proof We adapt the arguments leading to Equation (6) in the proof of Theorem 3. We begin by noticing that, if arm k is selected, then at least one of the three following equations holds:

$$B_{k^*,T_{k^*}(t-1),t} \leq \mu^*,$$
$$\hat{X}_{k,t} \geq \mu_k + \sqrt{\frac{f_k(t)}{T_k(t-1)}},$$
$$T_k(t-1) < \frac{4f_k(n)}{\Delta_k^2},$$

and the rest follows straightforwardly. ■

We are now able to give sufficient conditions on the f_k for UCB to be Hannan consistent.

Proposition 12 *If $f_k(n) = o(n)$ for all $k \in \{1,\dots,K\}$, and if there exists $\gamma > \frac{1}{2}$ and $N \geq 1$ such that $f_k(n) \geq \gamma \log\log n$ for all $k \in \{1,\dots,K\}$ and for any $n \geq N$, then* UCB *is Hannan consistent.*

Proof Fix an index k of a suboptimal arm and choose $\beta \in (0,1)$ such that $2\beta\gamma > 1$. By means of Lemma 11, one has for n large enough:

$$\mathbb{E}_\theta[T_k(n)] \leq u + 2\sum_{t=u+1}^{n} \left(1 + \frac{\log t}{\log(\frac{1}{\beta})}\right) e^{-2\beta\gamma \log\log t},$$

where $u = \left\lceil \frac{4f_k(n)}{\Delta_k} \right\rceil$.
Consequently, we have:

$$\mathbb{E}_\theta[T_k(n)] \leq u + 2\sum_{t=2}^{n} \left(\frac{1}{(\log t)^{2\beta\gamma}} + \frac{1}{\log(\frac{1}{\beta})}\frac{1}{(\log t)^{2\beta\gamma-1}}\right). \tag{11}$$

Sums of the form $\sum_{t=2}^{n} \frac{1}{(\log t)^c}$ with $c > 0$ are equivalent to $\frac{n}{(\log n)^c}$ as $n \to +\infty$. Indeed, on the one hand we have

$$\sum_{t=3}^{n} \frac{1}{(\log t)^c} \leq \int_2^n \frac{dx}{(\log x)^c} \leq \sum_{t=2}^{n} \frac{1}{(\log t)^c},$$

so that $\sum_{t=2}^{n} \frac{1}{(\log t)^c} \sim \int_2^n \frac{dx}{(\log x)^c}$. On the other hand, one can write

$$\int_2^n \frac{dx}{(\log x)^c} = \left[\frac{x}{(\log x)^c}\right]_2^n + c\int_2^n \frac{dx}{(\log x)^{c+1}}.$$

As both integrals are divergent we have $\int_2^n \frac{dx}{(\log x)^c} = o\left(\int_2^n \frac{dx}{(\log x)^{c+1}}\right)$, so that $\int_2^n \frac{dx}{(\log x)^c} \sim \frac{n}{(\log n)^c}$.

Now, by means of Equation (11), there exists $C > 0$ such that

$$\mathbb{E}_\theta[T_k(n)] \leq \left\lceil \frac{4f_k(n)}{\Delta} \right\rceil + \frac{Cn}{(\log n)^{2\beta\gamma-1}},$$

and this proves Hannan consistency. ■

The fact that there is no logarithmic lower bound then comes from the following proposition (which is a straightforward adaptation of Propostion 1).

Proposition 13 *Let $0 \leq b < a \leq 1$ and $n \geq 1$. For $\theta = (\delta_a, \delta_b)$, the random variable $T_2(n)$ is uniformly upper bounded by $\frac{f_2(n)}{\Delta^2} + 1$. Consequently, the expected regret of* ucb *is upper bounded by $\frac{f_2(n)}{\Delta} + 1$.*

Then, if $f_1(n) = f_2(n) = \log\log n$, ucb is Hannan consistent and the expected regret is of order $\log\log n$ in all environments of the form (δ_a, δ_b). Hence the conclusion on the non-existence of logarithmic lower bounds.

References

R. Agrawal. Sample mean based index policies with o(log n) regret for the multi-armed bandit problem. Advances in Applied Mathematics, 27:1054–1078, 1995.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In Second international symposium on information theory, volume 1, pages 267–281. Springer Verlag, 1973.

J.-Y. Audibert, R. Munos, and C. Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. Theoretical Computer Science, 410(19):1876–1902, 2009.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. Mach. Learn., 47(2-3):235–256, 2002.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The nonstochastic multiarmed bandit problem. SIAM Journal on Computing, 32(1):48–77, 2003.

D. Bergemann and J. Valimaki. Bandit problems. 2008. In The New Palgrave Dictionary of Economics, 2nd ed. Macmillan Press.

S. Bubeck. Bandits Games and Clustering Foundations. PhD thesis, Université Lille 1, France, 2010.

S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvari. Online optimization in X-armed bandits. In Advances in Neural Information Processing Systems 21, pages 201–208. 2009.

A.N. Burnetas and M.N. Katehakis. Optimal adaptive policies for sequential allocation problems. Advances in Applied Mathematics, 17(2):122–142, 1996.

N. Cesa-Bianchi and G. Lugosi. Prediction, learning, and games. Cambridge Univ Pr, 2006.

P.A. Coquelin and R. Munos. Bandit algorithms for tree search. In Uncertainty in Artificial Intelligence, 2007.

S. Gelly and Y. Wang. Exploration exploitation in go: UCT for Monte-Carlo go. In Online trading between exploration and exploitation Workshop, Twentieth Annual Conference on Neural Information Processing Systems (NIPS 2006), 2006.

J. Honda and A. Takemura. An asymptotically optimal bandit algorithm for bounded support models. In Proceedings of the Twenty-Third Annual Conference on Learning Theory (COLT), 2010.

R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In Proceedings of the 40th annual ACM symposium on Theory of computing, pages 681–690, 2008.

R. D. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In Advances in Neural Information Processing Systems 17, pages 697–704. 2005.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics, 6:4–22, 1985.

D. Lamberton, G. Pagès, and P. Tarrès. When can the two-armed bandit algorithm be trusted? Annals of Applied Probability, 14(3):1424–1454, 2004.

C.L. Mallows. Some comments on cp. Technometrics, pages 661–675, 1973.

H. Robbins. Some aspects of the sequential design of experiments. Bulletin of the American Mathematics Society, 58:527–535, 1952.

A. Salomon and J.Y. Audibert. Deviations of stochastic bandit regret. In Algorithmic Learning Theory, pages 159–173. Springer, 2011.

G. Schwarz. Estimating the dimension of a model. The annals of statistics, 6(2):461–464, 1978.

W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 25(3/4):285–294, 1933.